Maximum Likelihood Estimation (MLE)

Magíster en Economía

Teoría Econométrica (Econometric Theory)

Prof. Luis Chancí

www.luischanci.com

Outline

1. From OLS to MLE
2. Likelihood, log-likelihood, score, and information
3. Two canonical examples
4. Asymptotic theory of MLE
5. Efficiency, invariance, and inference
6. Looking ahead: limited dependent variable models

> **Main idea: MLE chooses the parameter vector that makes the observed sample look most plausible under a specified probabilistic model.**

1. From OLS to MLE

Why Another Estimation Method?

In OLS, we estimate parameters by minimizing a quadratic loss:

$$\hat{\beta}_{OLS} = \arg\min_{\beta} (y - X\beta)'(y - X\beta).$$

MLE also solves an optimization problem, but from a different angle:

> **Instead of minimizing a loss, MLE chooses the parameter vector that makes the observed sample most likely under a fully specified probability model.**

This makes MLE both:

- an **estimation method**
- a **modeling framework**

Connection with OLS

There is a direct bridge from OLS to MLE.

In the Gaussian linear regression model,

$$y \mid X \sim \mathcal{N}(X\beta, \sigma^2 I_n),$$

the MLE for β is exactly

$$\hat{\beta}_{MLE} = \hat{\beta}_{OLS}.$$

> **So OLS is not outside the likelihood framework. It is a special case of MLE under normal disturbances.**

This is one reason MLE is a natural continuation of the OLS chapter.

Strengths and Cost of MLE

Relative to OLS or GMM, MLE has a clear advantage:

- it often delivers estimators with strong large-sample properties
- it gives a unified framework for estimation and inference
- it is the natural tool in many nonlinear models

But there is also a cost:

> **MLE requires stronger structure because we must specify a probability model for the data.**

That is, we move from moment conditions to a full density or probability mass function.

2. The Likelihood Framework

From Probability to Likelihood

Suppose we observe

$$\{w_i\}_{i=1}^{n},$$

where each w_i has density or pmf

$$f(w_i;\theta), \qquad \theta \in \Theta \subset \mathbb{R}^k.$$

If the observations are independent, then the joint density is

$$f(w_1,\ldots,w_n;\theta) = \prod_{i=1}^{n} f(w_i;\theta).$$

> **Definition — Likelihood Function**
>
> Given the observed sample, the likelihood function is $\mathcal{L}_n(\theta) = \prod_{i=1}^{n} f(w_i;\theta)$

Notice that is viewed as a function of θ.

The Maximum Likelihood Estimator

Definition — MLE

The Maximum Likelihood Estimator is

$$\hat{\theta}_{MLE} = \arg\max_{\theta \in \Theta} \mathcal{L}_n(\theta).$$

Intuition: among all parameter values, choose the one under which the observed sample looks most plausible.

Different values of θ imply different probability models for the data. MLE picks the best-fitting one.

Why Use the Log-Likelihood?

In practice, we maximize

$$\ell_n(\theta) = \ln \mathcal{L}_n(\theta) = \sum_{i=1}^{n} \ln f(w_i; \theta).$$

Why?

- the logarithm is strictly increasing, so the maximizer does not change
- products become sums
- derivatives and numerical optimization become much easier

Definition

$$\hat{\theta}_{MLE} = \arg\max_{\theta \in \Theta} \ell_n(\theta).$$

Score, Hessian, and Fisher Information

Definition — Score and Hessian

The **score** is the gradient of the log-likelihood:

$$s_n(\theta) = \frac{\partial \ell_n(\theta)}{\partial \theta} = \sum_{i=1}^{n} s_i(\theta).$$

The **Hessian** is the matrix of second derivatives:

$$H_n(\theta) = \frac{\partial^2 \ell_n(\theta)}{\partial \theta \, \partial \theta'} = \sum_{i=1}^{n} H_i(\theta).$$

At an interior optimum: $s_n(\hat{\theta}) = 0$. This is the likelihood analogue of the OLS normal equations.

Fisher Information

Definition — Fisher Information

The Fisher information in one observation is

$$\mathcal{I}(\theta) = \mathbb{E}[\boldsymbol{s}_i(\theta)\boldsymbol{s}_i(\theta)'].$$

Under standard regularity conditions,

$$\mathcal{I}(\theta) = -\mathbb{E}[\boldsymbol{H}_i(\theta)].$$

If the log-likelihood is sharply curved around the truth, the data are very informative about θ.

If the likelihood is flat, the data are less informative.

3. Canonical Examples

Example 1: Mean of a Normal Distribution

Suppose

$$z_i \overset{iid}{\sim} \mathcal{N}(\mu, 1), \qquad i = 1, \ldots, n.$$

The density is

$$f(z_i; \mu) = \frac{1}{\sqrt{2\pi}} \exp\left\{ -\frac{(z_i - \mu)^2}{2} \right\}.$$

Hence the log-likelihood is

$$\ell_n(\mu) = -\frac{n}{2}\ln(2\pi) - \frac{1}{2}\sum_{i=1}^{n}(z_i - \mu)^2.$$

The score is

$$s_n(\mu) = \sum_{i=1}^{n}(z_i - \mu).$$

Example 1: Solving the FOC

Set the score equal to zero:

$$\sum_{i=1}^{n}(z_i - \hat{\mu}) = 0.$$

Then $\hat{\mu}_{MLE} = \bar{z}$, and the Hessian is $H_n(\mu) = -n$, so the information in one observation is $\mathcal{I}(\mu) = 1$.

> **This simple example already shows the full MLE logic:**
>
> **specify a density $\rightarrow$ write the log-likelihood $\rightarrow$ compute the score $\rightarrow$ solve the FOC.**

Example 2: The Normal Linear Regression Model

Now consider

$$y = X\beta + u, \qquad u \mid X \sim \mathcal{N}(0, \sigma^2 I_n).$$

Then

$$y \mid X \sim \mathcal{N}(X\beta, \sigma^2 I_n),$$

and the log-likelihood is

$$\ell_n(\beta, \sigma^2) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}(y - X\beta)'(y - X\beta).$$

> **For fixed σ^2, maximizing the log-likelihood with respect to β is equivalent to minimizing the sum of squared residuals.**

Example 2: MLE and OLS Coincide

Differentiate with respect to β:

$$\frac{\partial \ell_n(\beta, \sigma^2)}{\partial \beta} = \frac{1}{\sigma^2} X'(y - X\beta).$$

Set equal to zero:

$$X'X\hat{\beta} = X'y.$$

Therefore,

$$\hat{\beta}_{MLE} = (X'X)^{-1}X'y = \hat{\beta}_{OLS}.$$

For the variance parameter,

$$\hat{\sigma}^2_{MLE} = \frac{(y - X\hat{\beta})'(y - X\hat{\beta})}{n}.$$

> **This differs from the usual unbiased OLS estimator of σ^2, which divides by $n - k$ rather than n.**

4. Asymptotic Theory of MLE

Regularity Conditions: Big Picture

For MLE asymptotics, we need:

- correct specification of the model
- identification of the true parameter
- smoothness of the log-likelihood
- nonsingularity of the information matrix
- LLN and CLT conditions for the score and Hessian

> **Regularity conditions for MLE**
>
> These conditions ensure that the expected log-likelihood is uniquely maximized at θ_0 and that derivatives behave well enough for Taylor expansions and probabilistic limits.

Consistency of MLE

Let

$$Q_n(\theta) = \frac{1}{n}\ell_n(\theta).$$

Under a suitable LLN,

$$Q_n(\theta) \xrightarrow{p} Q(\theta) = \mathbb{E}[\ell_i(\theta)]$$

uniformly over θ.

If the population criterion $Q(\theta)$ is uniquely maximized at θ_0, then

$$\hat{\theta}_{MLE} \xrightarrow{p} \theta_0.$$

Consistency of MLE

Consistency of MLE

Under standard regularity conditions,

$$\hat{\theta}_{MLE} \xrightarrow{p} \theta_0 .$$

MLE is an extremum estimator: consistency comes from the sample criterion converging to a population criterion with a unique maximizer.

Asymptotic Normality: Step 1

The score satisfies

$$s_n(\hat{\theta}) = 0.$$

Expand around θ_0:

$$0 = s_n(\theta_0) + H_n(\tilde{\theta})(\hat{\theta} - \theta_0),$$

where $\tilde{\theta}$ lies between $\hat{\theta}$ and θ_0.

Rearranging,

$$\sqrt{n}(\hat{\theta} - \theta_0) = -\left[\frac{1}{n} H_n(\tilde{\theta})\right]^{-1} \left[\frac{1}{\sqrt{n}} s_n(\theta_0)\right].$$

This isolates the two key objects: the score and the Hessian

Asymptotic Normality: Step 2

Because

$$s_n(\theta_0) = \sum_{i=1}^{n} s_i(\theta_0),$$

and

$$\mathbb{E}[s_i(\theta_0)] = 0,$$

the CLT gives

$$\frac{1}{\sqrt{n}} s_n(\theta_0) \xrightarrow{d} \mathcal{N}(0, \mathcal{I}(\theta_0)).$$

Since $\hat{\theta} \xrightarrow{p} \theta_0$, we also have $\tilde{\theta} \xrightarrow{p} \theta_0$, and by LLN

$$-\frac{1}{n} H_n(\tilde{\theta}) \xrightarrow{p} \mathcal{I}(\theta_0).$$

Asymptotic Normality: Final Result

By Slutsky's theorem,

$$\sqrt{n}(\hat{\theta}_{MLE} - \theta_0) \xrightarrow{d} \mathcal{N}\left(0, \mathcal{I}(\theta_0)^{-1}\right).$$

Asymptotic normality of MLE

Under standard regularity conditions,

$$\sqrt{n}(\hat{\theta}_{MLE} - \theta_0) \xrightarrow{d} \mathcal{N}\left(0, \mathcal{I}(\theta_0)^{-1}\right).$$

This is the foundation for standard errors, confidence intervals, and large-sample hypothesis tests in MLE.

Variance Estimation in Practice

The asymptotic variance depends on the unknown information matrix, so we need to estimate it.

Three common approaches are:

- **Observed information**

$$\widehat{\mathrm{Var}}_H(\hat{\theta}) = [-H_n(\hat{\theta})]^{-1}$$

- **Outer product of gradients (OPG)**

$$\widehat{\mathrm{Var}}_{OPG}(\hat{\theta}) = \left(\sum_{i=1}^{n} \boldsymbol{s}_i(\hat{\theta}) \boldsymbol{s}_i(\hat{\theta})' \right)^{-1}$$

- **Sandwich form under misspecification**

$$\widehat{\mathrm{Var}}_{sand}(\hat{\theta}) = \frac{1}{n} A_n^{-1} B_n A_n^{-1}$$

where

$$A_n = -\frac{1}{n} H_n(\hat{\theta}), \qquad B_n = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{s}_i(\hat{\theta}) \boldsymbol{s}_i(\hat{\theta})'.$$

5. Efficiency, Invariance, and Inference

The Cramér–Rao Lower Bound

For an unbiased scalar estimator $\tilde{\theta}$,

$$\mathbb{V}(\tilde{\theta}) \geq \frac{1}{\mathcal{I}_n(\theta_0)},$$

where

$$\mathcal{I}_n(\theta_0) = n\mathcal{I}(\theta_0).$$

> **Intuition: the more sharply the likelihood responds to changes in θ, the more informative the data are, and the smaller the lower bound on variance.**

This is the Cramér–Rao lower bound.

Why MLE Is Called Efficient

MLE is not necessarily unbiased in finite samples, so the finite-sample Cramér–Rao result should be interpreted carefully.

What is true is that, under correct specification,

$$\sqrt{n}(\hat{\theta}_{MLE} - \theta_0) \xrightarrow{d} \mathcal{N}(0, \mathcal{I}(\theta_0)^{-1}),$$

and this is the smallest asymptotic covariance matrix available among regular estimators.

Asymptotic efficiency of MLE

Under correct specification and standard regularity conditions, MLE attains the asymptotic information bound.

The Invariance Property

Invariance of MLE

If $\gamma = g(\theta)$ and $\hat{\theta}_{MLE}$ is the MLE of θ, then

$$\hat{\gamma}_{MLE} = g(\hat{\theta}_{MLE}).$$

Example:

if the MLE of σ^2 is $\hat{\sigma}^2_{MLE}$, then the MLE of σ is simply

$$\hat{\sigma}_{MLE} = \sqrt{\hat{\sigma}^2_{MLE}}.$$

> **You do not need to solve a new optimization problem for every smooth transformation of the parameters.**

Wald Tests

Suppose

$$H_0 : R\theta = r.$$

The Wald statistic is

$$W = (R\hat{\theta} - r)' \left[R\widehat{\mathrm{Var}}(\hat{\theta}) R' \right]^{-1} (R\hat{\theta} - r).$$

Under H_0,

$$W \xrightarrow{d} \chi^2_q.$$

> **The Wald test asks whether the unrestricted estimate lies far from the null restriction once that distance is scaled by estimation uncertainty.**

Likelihood Ratio and Score Tests

Let $\hat{\theta}_u$ be the unrestricted MLE and $\tilde{\theta}_r$ the restricted MLE.

Likelihood Ratio test

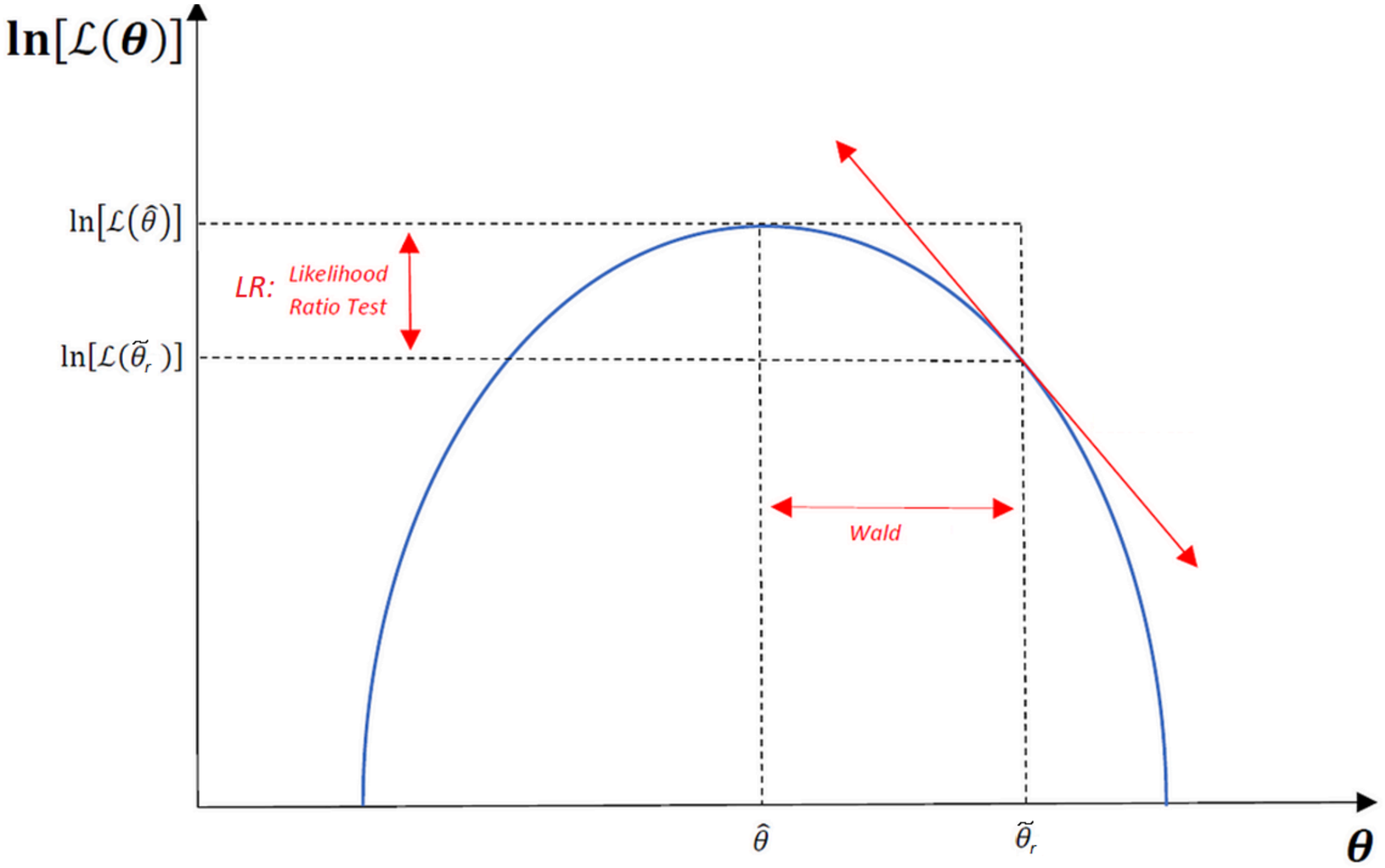
$$LR = 2\left[\ell_n(\hat{\theta}_u) - \ell_n(\tilde{\theta}_r)\right] \xrightarrow{d} \chi^2_q.$$

Score / LM test

$$LM = \boldsymbol{s}_n(\tilde{\theta}_r)' \left[n\widehat{\mathcal{I}}(\tilde{\theta}_r)\right]^{-1} \boldsymbol{s}_n(\tilde{\theta}_r) \xrightarrow{d} \chi^2_q.$$

> **All three tests are asymptotically equivalent under the null, but they differ in how much estimation they require.**

Geometric Intuition: Wald vs. LR



- **Wald**: measures how far the unrestricted estimate is from the null set
- **LR**: measures how much the fit worsens when the restriction is imposed

6. Looking Ahead

Why MLE Matters Beyond Gaussian Regression

The real power of MLE appears when OLS is no longer natural.

Examples:

- binary outcomes $\rightarrow$ **logit** and **probit**
- count outcomes $\rightarrow$ **Poisson**
- censored outcomes $\rightarrow$ **Tobit**
- duration data $\rightarrow$ **hazard / survival models**

> **The logic remains the same: specify a probabilistic model, write the likelihood, optimize it, and use information-based asymptotic theory for inference.**

Summary

1. MLE chooses the parameter vector that makes the observed sample most plausible under a specified probability model.
2. The score, Hessian, and Fisher information organize both estimation and inference.
3. In the Gaussian linear model, MLE connects directly back to OLS.
4. Under standard regularity conditions,

$$\sqrt{n}(\hat{\theta}_{MLE} - \theta_0) \xrightarrow{d} \mathcal{N}\left(0, \mathcal{I}(\theta_0)^{-1}\right).$$

5. Wald, LR, and Score tests emerge naturally in the likelihood framework.

Cierre

¿Preguntas?

O vía E-mail:

luis.chanci@usach.cl

luischanci@santotomas.cl